CVPR SEATTLE WASHINGTON
JUNE 16-18 2020

EventCap: High-Speed Human Motion Capture using an Event Camera

Lan XU 许岚

Hong Kong University of Science and Technology

2020/06/11

max planck institut informatik
Graphics Vision & Video
SIC Saarland Informatics Campus
香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY
清華大学 Tsinghua University

Background

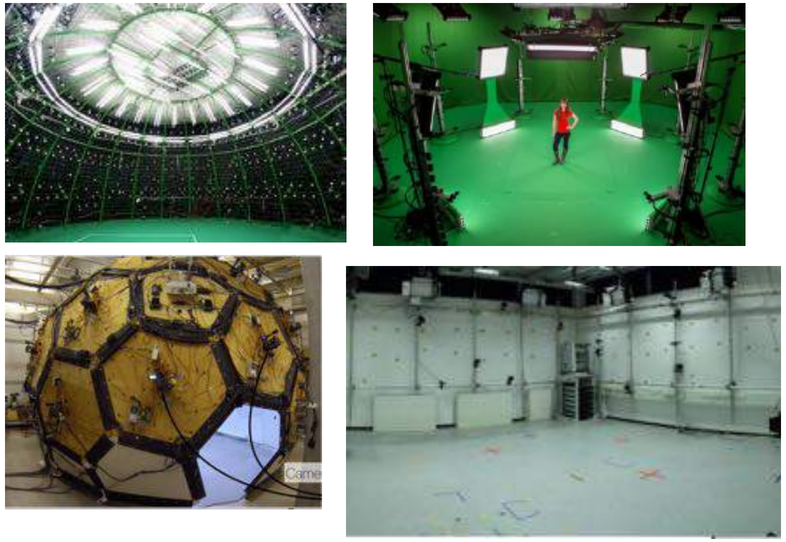
1. Background

❑ Previous MoCap systems

1st Generation

Marker-based MoCap :
- Only reconstruct makers
- Intrusive, restricted clothing
- Not ready for daily usage

2nd Generation



High-end marker-less system:
- Many, many cameras
- Green background & fixed space
- Tedious synchronization, calibration

3rd Generation

Convenient Capture
- Handheld or single-view
- Consumer-level
- Still fixed captured volume

Technological Trend: Realtime, convenient and high quality
4D human reconstruction is critical

1. Background

❑ Bottleneck of high-speed human MoCap

UnstructuredFusion

RobustFusion

MonoPerfCap

LiveCap

- high speed motion analysis is rare
- RGB/RGBD: good lighting for high frame rates
- Throughput: a VGA RGB stream at 1000 fps for 60 s → 51.5 GB !!!

1. Background

❑ Bottleneck of high-speed human MoCap

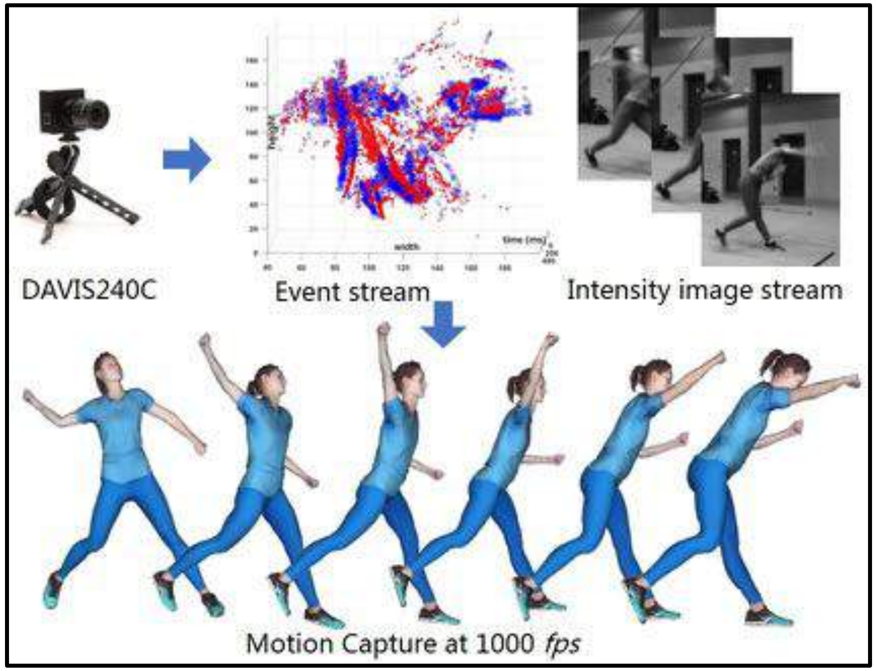
- high speed motion analysis is rare
- RGB/RGBD: ... lighting for high frame rates
- Throughput: a VGA RGB stream at 100 fps for 60 s -> 51.5 GB

Could we liberates these constraints ?

EventCap: use new device !!!

Key idea

2. Human Modelling: EventCap



Basic idea

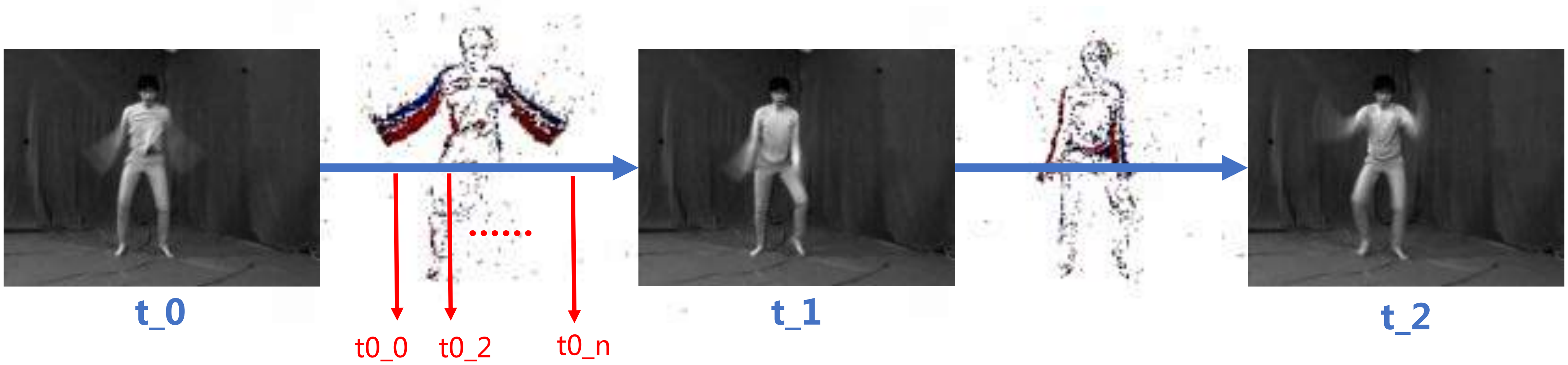
- Capturing **high-speed** human motions at **1000 fps**

Benefits:

- High temporal resolution, HDR (140 dB), low data bandwidth

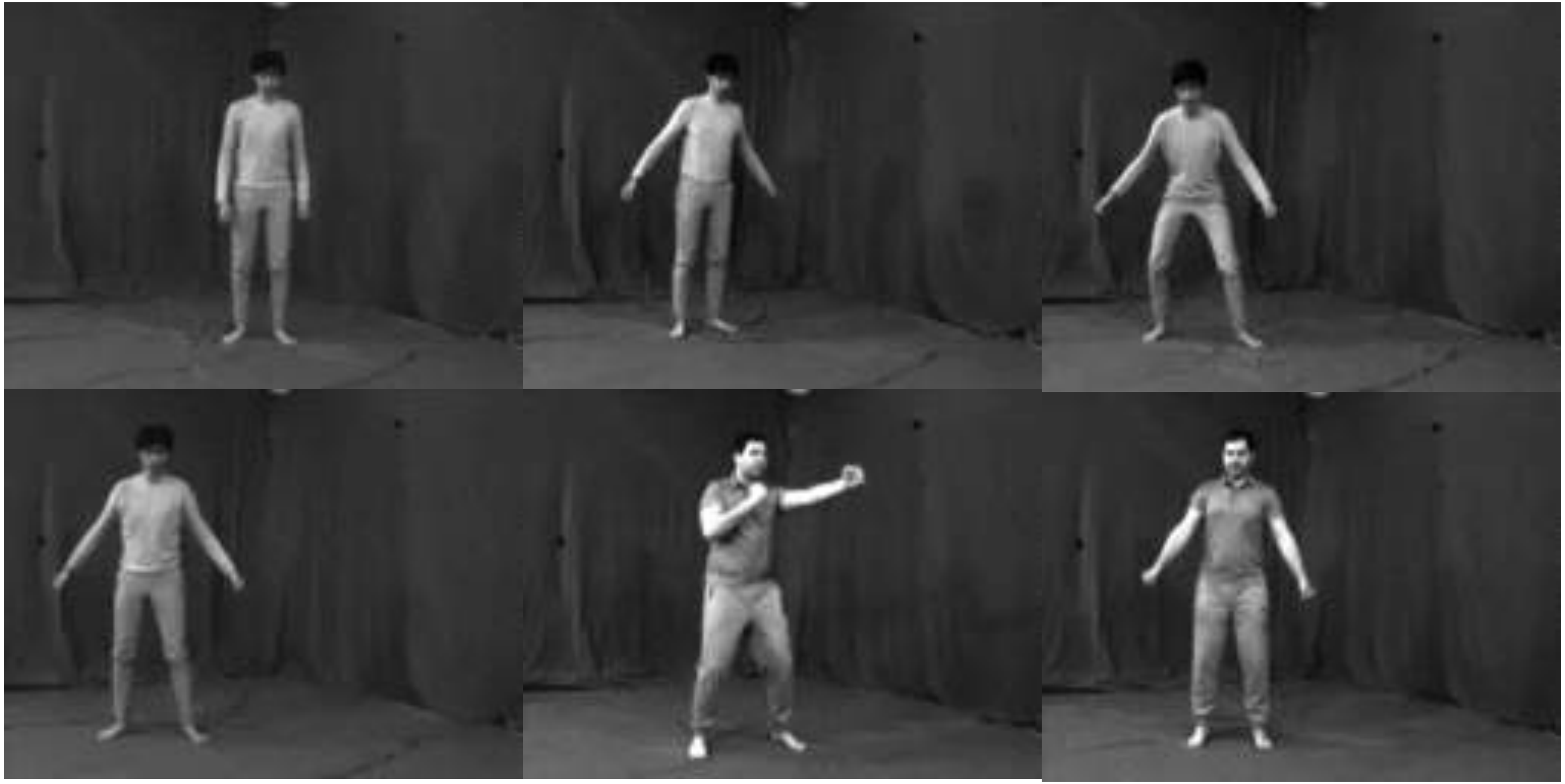
Challenges:

- Images & events: **unstructured** temporal information
- Severe image blur

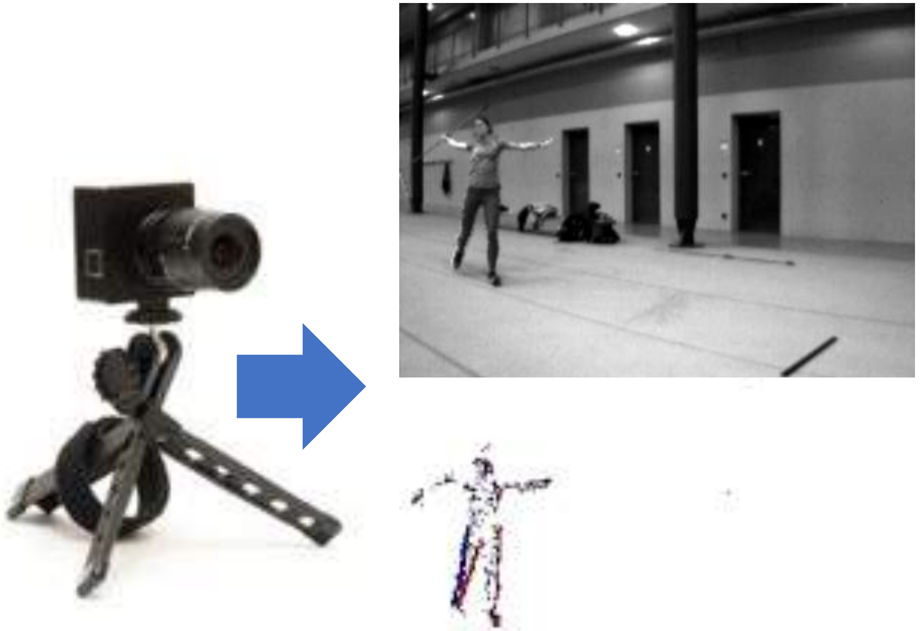


2. Human Modelling: EventCap

- **High-speed human motions**



2. Human Modelling: EventCap



V.S.

Event camera: DAVIS240C

High speed camera: Sony RX0

Only 3.4% data bandwidth

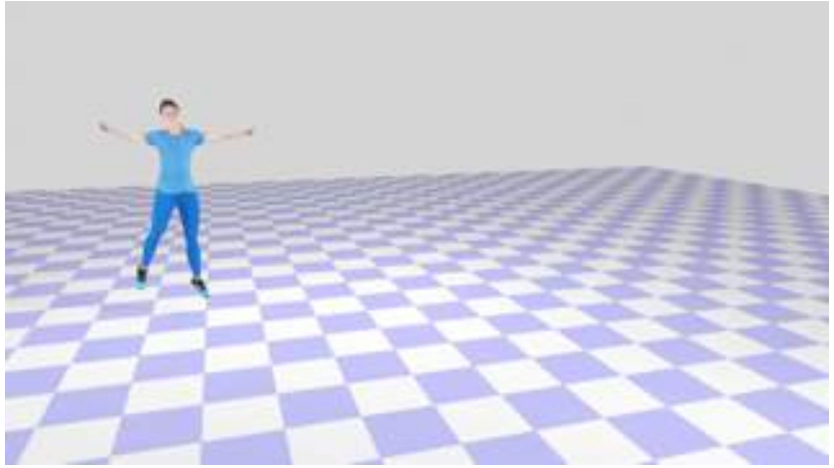
2. Human Modelling: EventCap

Reconstruction results for sports analysis

Low FPS image

Event polarity

Reference view in Sony Camera

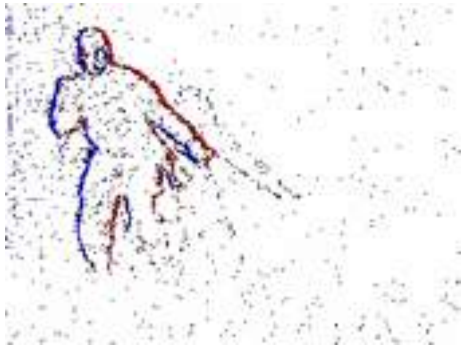


2. Human Modelling: EventCap

❑ Results of capturing a Ninja in the dark

- Thanks to the high dynamic range (140 dB) of the event camera

Low FPS image



Event polarity

(Original images)

(Gamma enhancement)

Reference view in Sony Camera

Method

3. Algorithm details

❑ Input of EventCap

Rigged Template

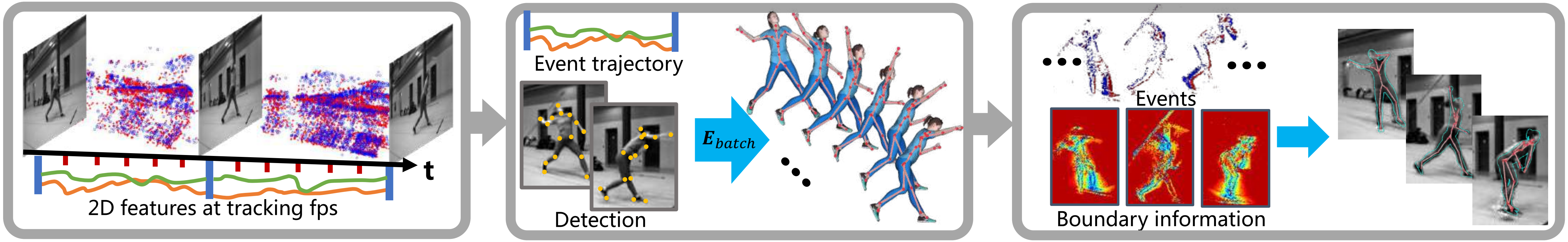
DAVIS 240C

Intensity image stream

Event stream

3. Algorithm details

❑ Framework



1. Event Trajectory Generation **2. Batch Optimization** **3. Event-based Pose Refinement**

3. Algorithm details

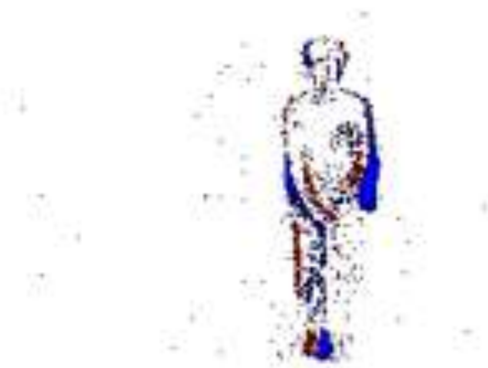
❑ Stage I: Event Trajectory Generation

3. Algorithm details

❑ Stage I: Event Trajectory Generation

Intensity image stream

+



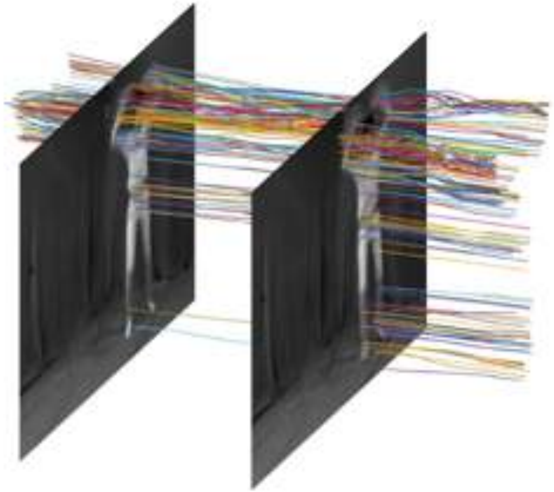
Event stream

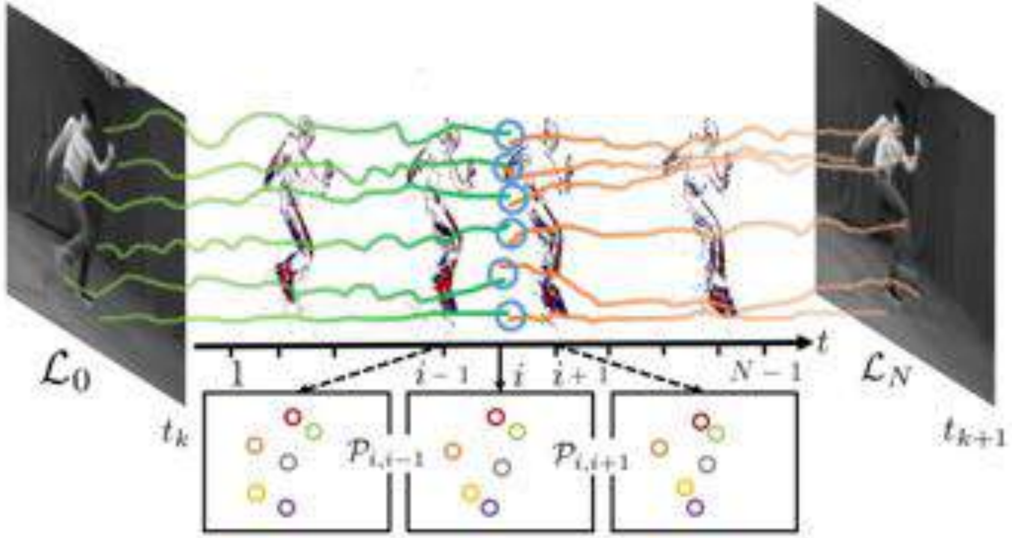
=

Event trajectories

3. Algorithm details

❑ Stage I: Event Trajectory Generation

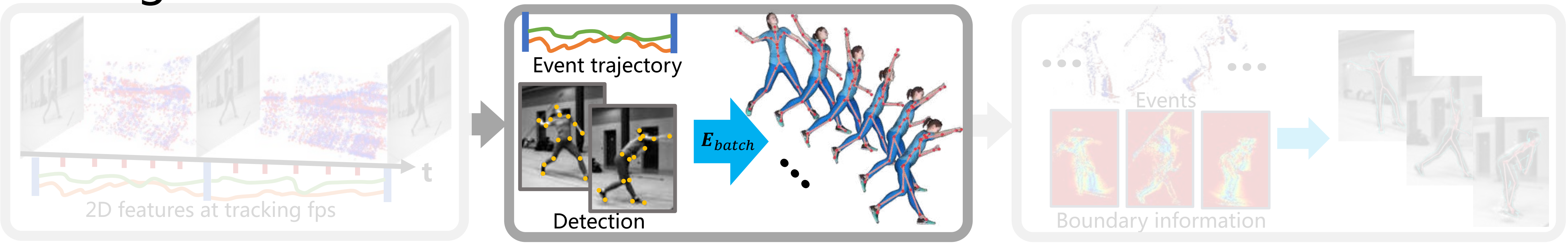




- 2D feature trajectory between adjacent images
- Forward & backward alignment
- Trajectory slicing → 2D correspondence pairs

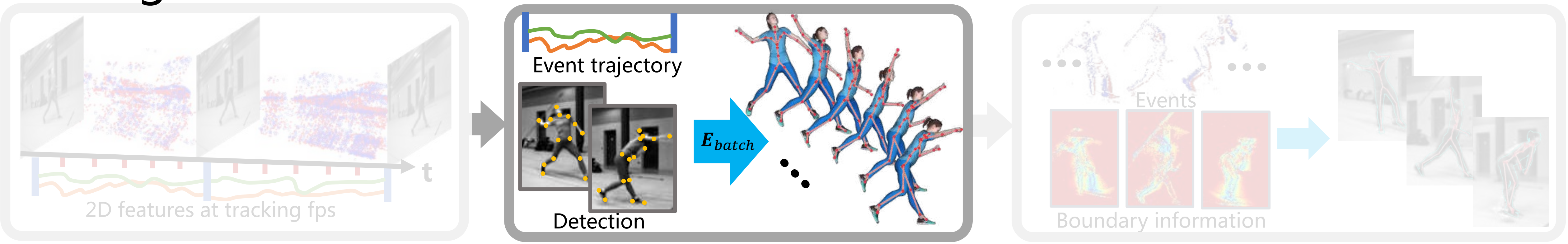
3. Algorithm details

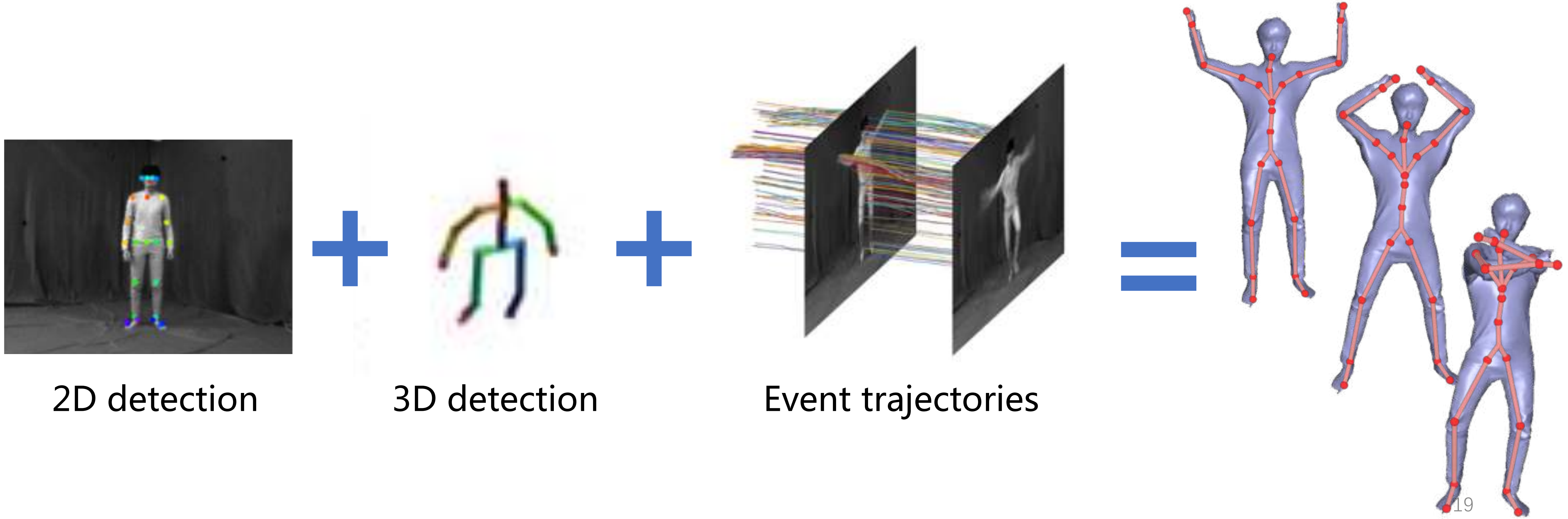
Stage II: Batch Optimization



3. Algorithm details

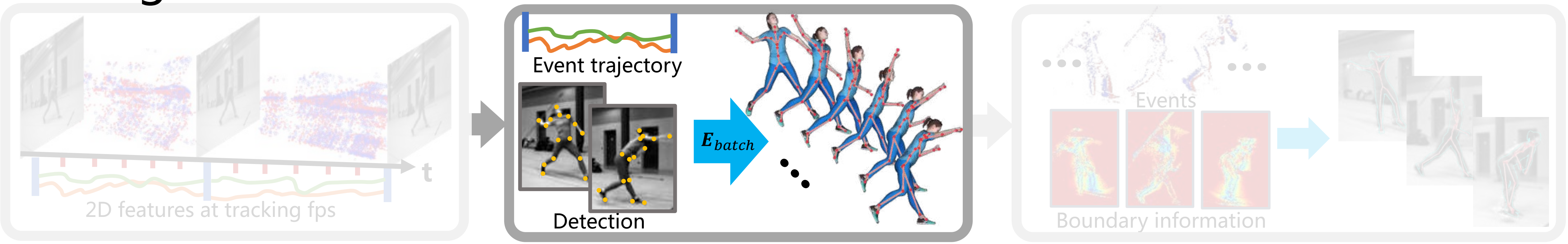
Stage II: Batch Optimization

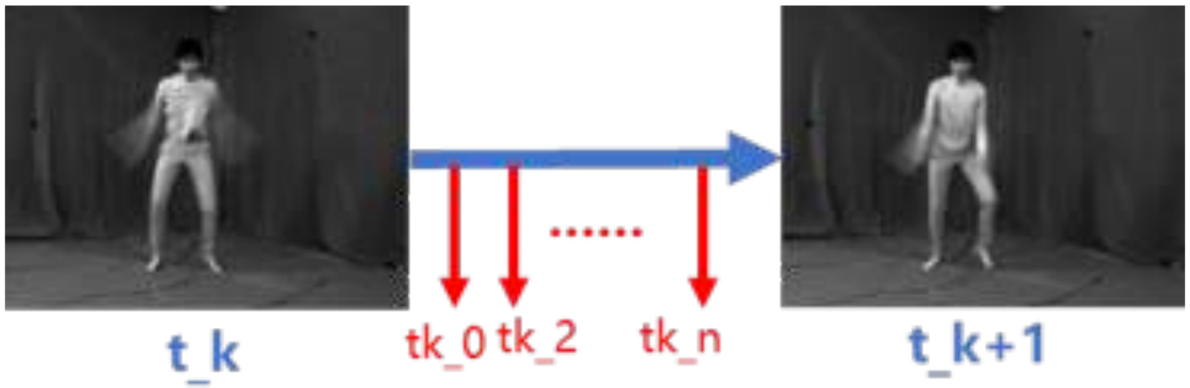




3. Algorithm details

Stage II: Batch Optimization





$$\mathbf{S} = \{S_f\}, f \in [0, N]$$

$$\mathbf{S}^* = \arg\min_{\mathbf{S}} \boldsymbol{E}_{\text{batch}}(\mathbf{S})$$

$$\text{s.t.} \quad \theta_{min} \le \theta_f \le \theta_{max}, \quad \forall f \in [0, N],$$

$$\boldsymbol{E}_{\text{batch}}(\mathbf{S}) = \lambda_{\text{adj}} \boldsymbol{E}_{\text{adj}} + \lambda_{\text{2D}} \boldsymbol{E}_{\text{2D}} + \lambda_{\text{3D}} \boldsymbol{E}_{\text{3D}} + \lambda_{\text{temp}} \boldsymbol{E}_{\text{temp}}.$$

$$\boldsymbol{E}_{\text{adj}}(\mathbf{S}) = \sum_{(i,j)\in\mathcal{C}} \sum_{h=1}^{H} \tau(p_{i,h}) \|\pi(v_{i,h}(S_j)) - p_{j,h}\|_2^2,$$

$$\boldsymbol{E}_{\text{2D}}(\mathbf{S}) = \sum_{f\in\{0,N\}} \sum_{l=1}^{N_J+4} \|\pi(J_l(S_f)) - \mathbf{P}_{f,l}^{2D}\|_2^2,$$

$$\boldsymbol{E}_{\text{3D}}(\mathbf{S}) = \sum_{f\in\{0,N\}} \sum_{l=1}^{N_J} \|J_l(S_f) - (\mathbf{P}_{f,l}^{3D} + \mathbf{t}')\|_2^2,$$

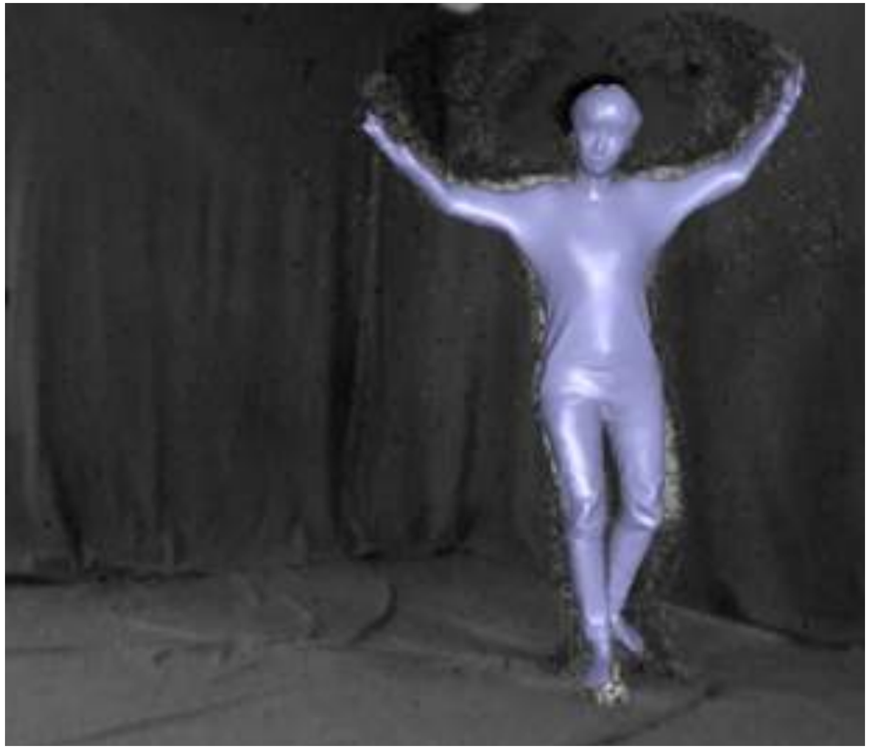
$$\boldsymbol{E}_{\text{temp}}(\mathbf{S}) = \sum_{\substack{(i,j)\in\mathcal{C} \\ i>j}} \sum_{l=1}^{N_J} \phi(l) \|J_l(S_i) - J_l(S_j)\|_2^2,$$

3. Algorithm details

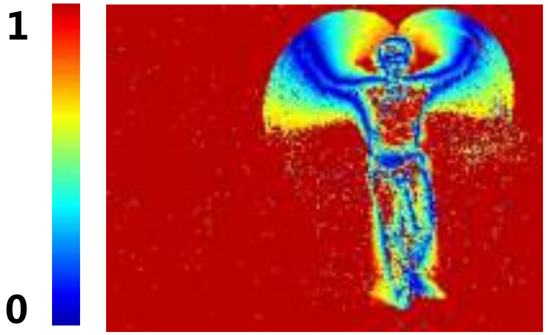
❑ Stage II: Event-based Pose Refinement

3. Algorithm details

Stage II: Event-based Pose Refinement



Results of Stage II



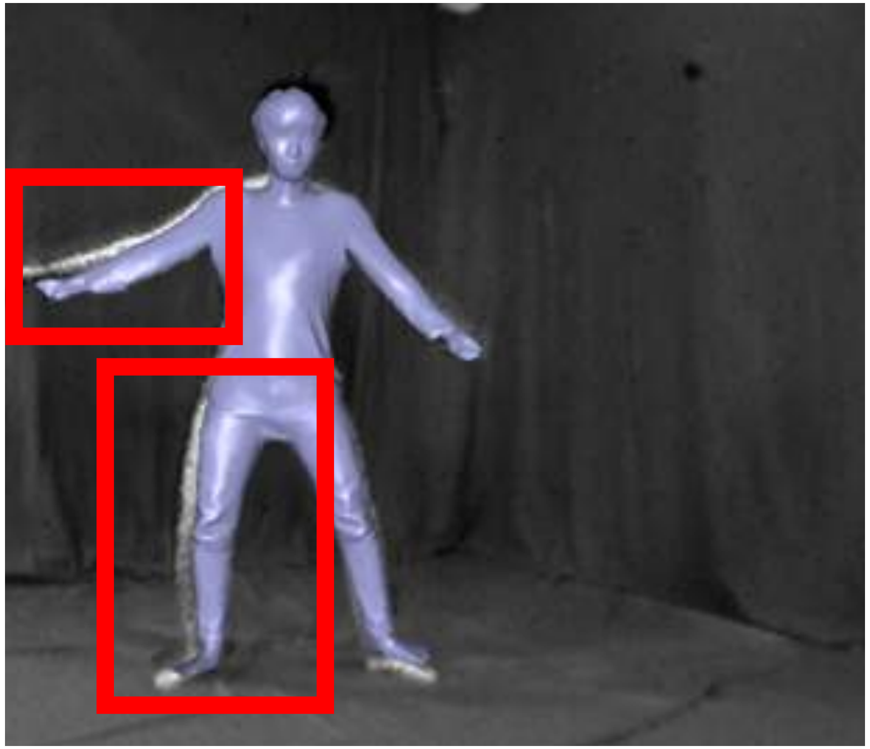
Normalized distance map

Event stream

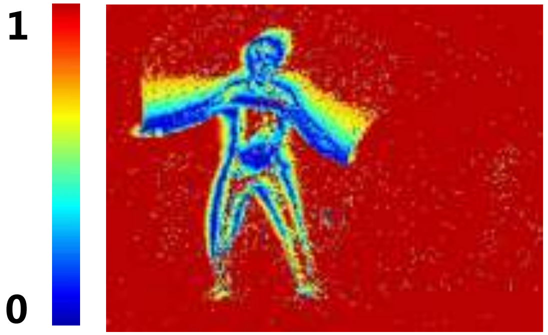
Our final results

3. Algorithm details

❑ Stage II: Event-based Pose Refinement

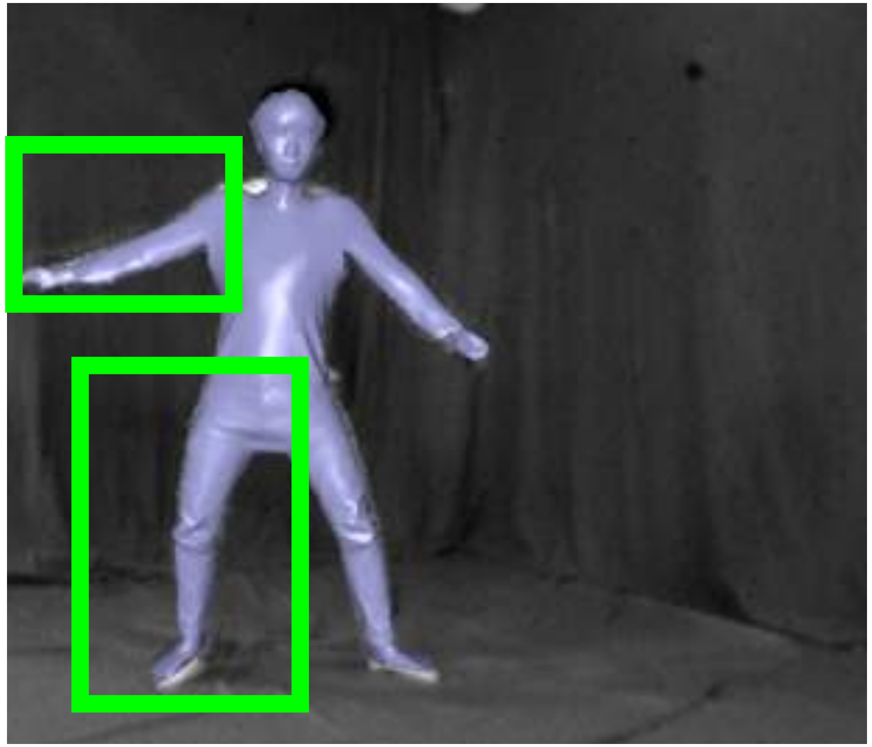


Results of Stage II



Normalized distance map

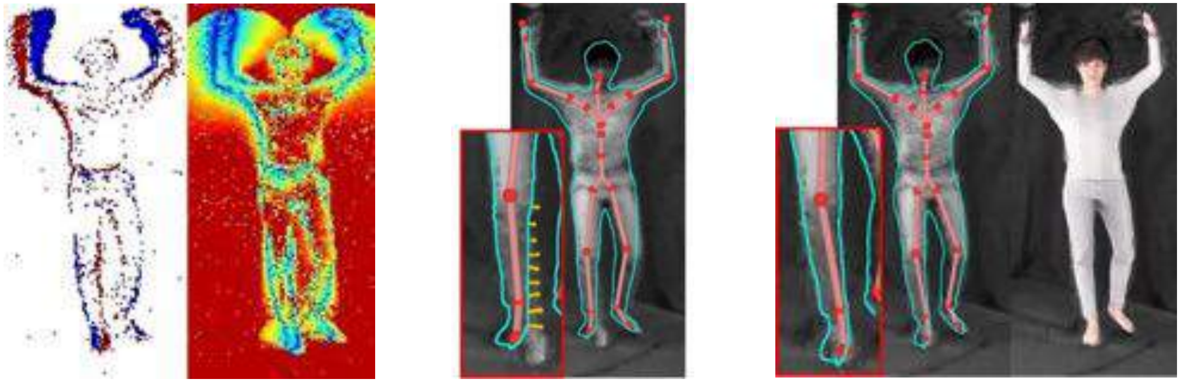
Event stream



Our final results

3. Algorithm details

Stage II: Event-based Pose Refinement



Distance map

Before refinement

After refinement

$$\boldsymbol{E}_{\text{refine}}(S_f) = \lambda_{\text{sil}}\boldsymbol{E}_{\text{sil}}(S_f) + \lambda_{\text{stab}}\boldsymbol{E}_{\text{stab}}(S_f).$$

$$\boldsymbol{E}_{\text{stab}}(S_f) = \sum_{l=1}^{N_J} \| J_l(S_f) - J_i(\hat{S}_f) \|_2^2,$$

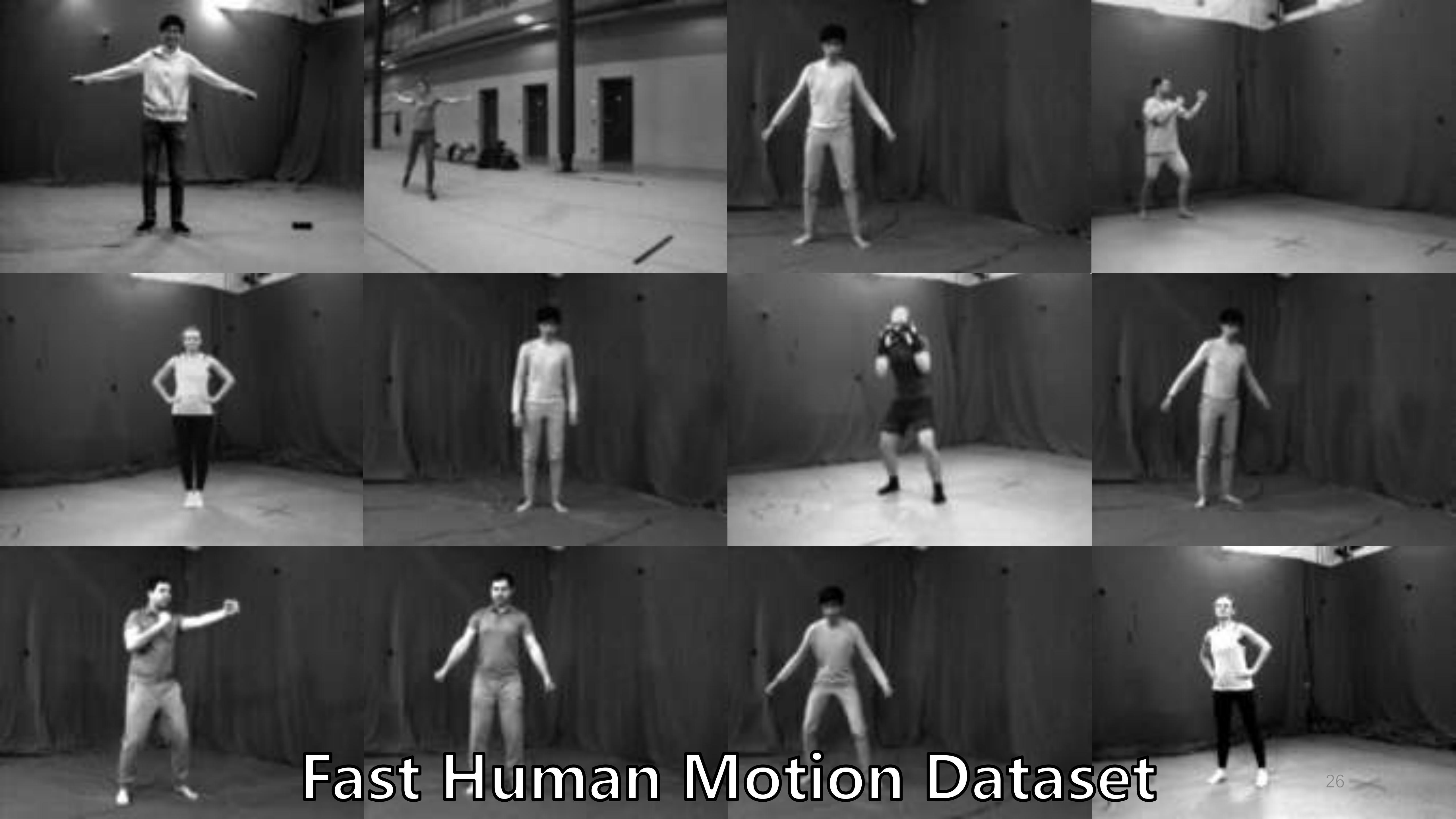
$$\boldsymbol{E}_{\text{sil}}(S_f) = \sum_{b \in \mathcal{B}} \| \mathbf{n}_b^{\mathrm{T}} (\pi(v_b(S_f) - u_b)) \|_2^2,$$

$$u_b = \arg\min_{u \in \mathcal{P}} \mathcal{D}(u, s_b).$$

$$\mathcal{D}(u, s_b) = \lambda_{dist} \| \mathcal{B}(t_f, u) \|_2^2 + \| u - s_h \|_2^2,$$

$$\mathcal{B}(t_f, u) = \mathbf{min}(\mathbf{abs}(t_u - t_f)/(t_N - t_0), 1),$$

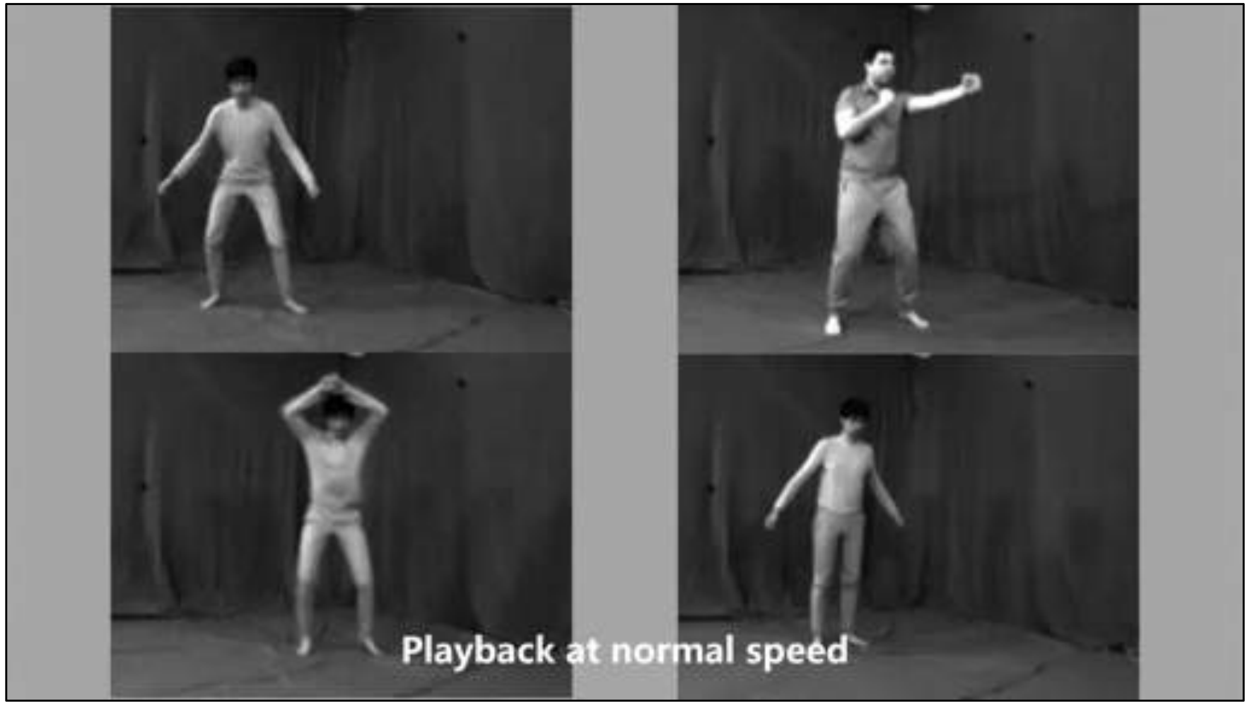
Results



Fast Human Motion Dataset

4. Results of EventCap

❑ Reconstruction results



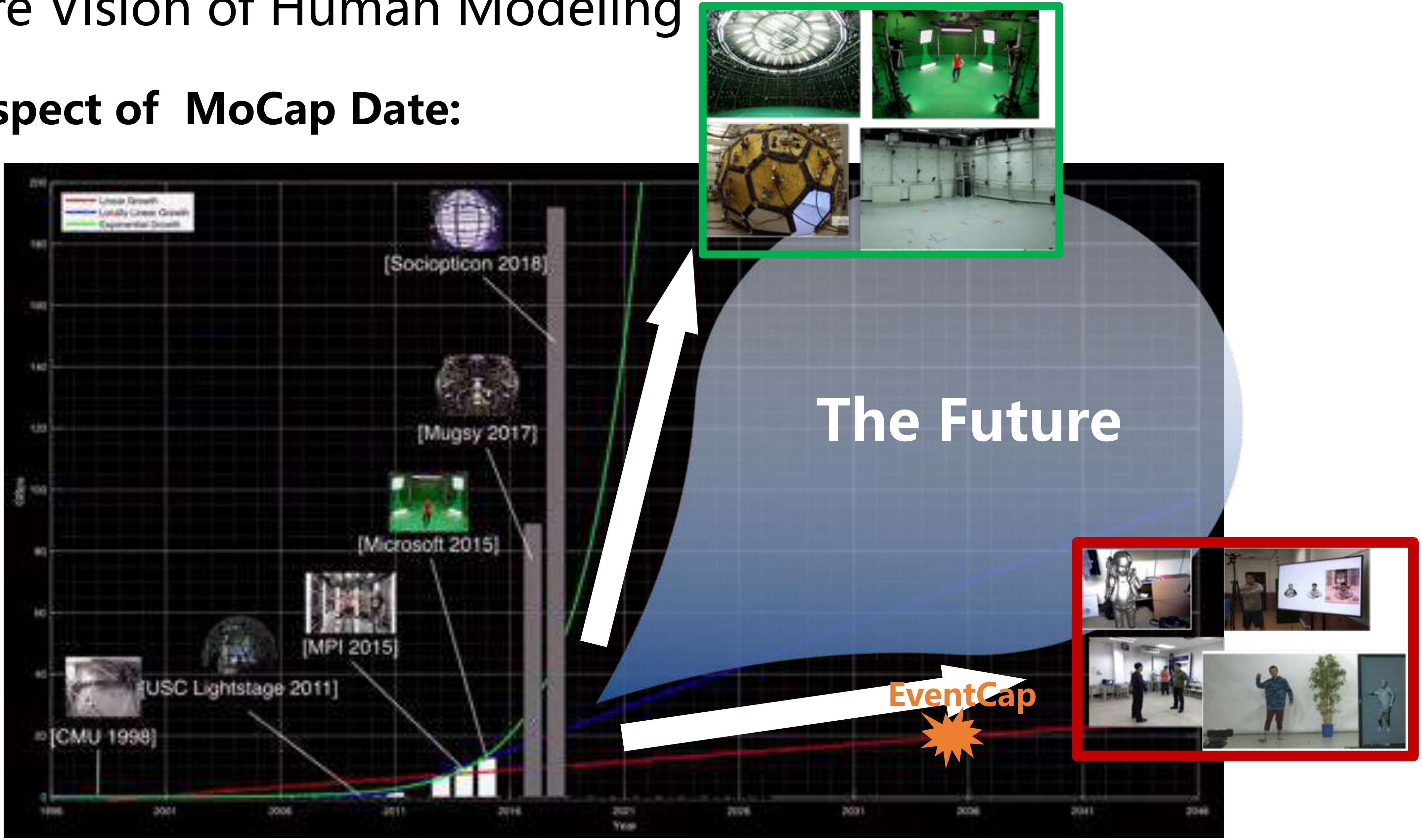
https://www.xu-lan.com/research.html

Summary

5. Future Vision of Human Modeling

- Aspect of MoCap Date:

Throughput (GBps)



Year

Figure from Prof. Yaser Sheikh, CMU

Thanks for your attention!

The End